A purely statistical bias in efficiency re-weightings - investigations in `PIDCalib`

Internal Note

Issue: 1
Revision: 0

Reference: LHCb-XX-XXXX
Created: Saturday 21st June, 2014
Last modified: Saturday 21st June, 2014

Prepared by: Stephen Ogilvy

Abstract

A purely statistical bias in efficiency re-weighting techniques is identified and demonstrated. This effect is common in all efficiency re-weighting procedures in which some calibration dataset is used to construct binned efficiencies. The magnitude of the bias is dependent on the magnitude of the efficiency of the selection being corrected for, on the statistical error from the binomial error on the bin efficiencies, and on the number of bins in a given schema. The effect of the bias is always to underestimate average re-weighted efficiencies. This bias was investigated for likely re-weighting procedures in `PIDCalib`. While in most use cases the calibration datasets are large enough to make the bias minimal, where calibration datasets are smaller, as is the case with the new proton samples, the effect becomes significant.

Document Status Sheet

1. Document Title: A purely statistical bias in efficiency re-weightings - investigations in `PIDCalib`			
2. Document Reference Number: LHCb-XX-XXXX			
3. Issue	**4. Revision**	**5. Date**	**6. Reason for change**
Draft	1	June 20, 2014	First version.

Contents

List of Figures

List of Tables

1 Introduction

Efficiency re-weighting procedures are commonly conducted on LHCb. Arguably the most prevalent is the PIDCalib technique. In this procedure, so-called "golden modes" are reconstructed and selected through decay kinematics alone, without the use of PID discrimination. The tracks from these modes are then used to construct local efficiencies by binning in variables on which the PID response is dependent (usually some combination of p, p_{T}, η, nTracks etc.). This is necessary as the kinematics of decay modes of interest are different to the calibration samples.

The systematic uncertainties in this procedure come from both increased binomial uncertainties due to the division of the dataset into bins with smaller populations than the whole, and from variations in PID cut efficiency across single kinematic bins. The second is well known to result in biases in local regions where the distributions of calibration and reference samples have very different kinematics. It is the intention of this document to demonstrate that the inflated binomial uncertainties, an effect which is known to result in an increased binomial uncertainty on the propagated reference sample efficiency, can also manifest in biases on the re-weighted PID efficiencies of reference samples. This bias is entirely statistical in nature, and is common to all instances of PIDCalib.

The severity of these effects are typically not large given the standard PIDCalib applications. Typical hadronic PID calibrations involve PID cut efficiencies of order 10^{-1}, with tens of millions of calibration tracks. As such, the binomial errors on calibration bins are low enough that these biases become negligible. Some particular cases where the calibration statistics or the PID cut efficiency are very low were identified where the bias might become non-negligible and investigated using toy MC. These were:

- Efficiency calibrations with the new protons samples (currently in development). The calibration dataset size is of order 10^5, and it was found that with such low calibration statistics the biases become significant.
- Evaluations of $K \rightarrow \mu$ mis-ID with tight cuts, as in the $B^0_s \rightarrow \mu^+\mu^-$ analysis where the mis-ID efficiency can be as low as 0.002 in some kinematic regions. It was found that the high number of K calibration statistics prevents the bias from becoming significant, even with such low mis-ID efficiencies.

In this document the bias is outlined and the toy study results presented.

2 A statistical bias

Consider a general calibration dataset of N_{t} tracks. We apply some PID selection ϵ to this population and are left with N_{p} tracks surviving the selection:

$$N_{\mathrm{p}} = N_{\mathrm{t}}\epsilon. \tag{1}$$

In the unrealistic case whereby the track kinematics are the same in the calibration and reference samples, this ϵ can be used to correct for the efficiency of a reference sample, with a known number of events passing PID selection k, to extract the number of reference tracks before selection, l - this l can be referred to as the adjusted track yield.

Realistically, the distributions of kinematics are different in the reference and calibration samples. The PID response depends on these, so we must use a re-weighting process to extract an efficiency for the reference sample. In these cases, the calibration dataset is binned in the kinematics, where local efficiencies in the kinematic space are derived. The efficiency for bin j is then given as:

$$\epsilon_{\mathrm{j}} = \frac{k_j}{l_j}. \tag{2}$$

To illustrate this bias it is easiest to consider binnings in variables on which the efficiency is not dependent, which unfolds biases from efficiency variation over individual efficiency bins. The considerations

nonetheless apply equally well to those variables on which the efficiency does depend. Now, in this particular example we consider a binning in a variable, x, which is uncorrelated with the PID efficiency. We bin the calibration data in x, apply some arbitrary PID selection and construct local efficiencies. Without efficiency variation in x, each bin efficiency will be roughly the same but of course statistical fluctuations will result in a distribution of bin efficiencies. Assuming a large enough number of bins, and a large enough number of calibration tracks populating each bin, the distribution of bin effiencies will be Gaussian-distributed about the effective efficiency for the integrated dataset. The width of the Gaussian will be roughly the average binomial error on the bin efficiencies.

Now say we take a sample of reference data, which has had the same PID selection applied to it. We assume the distribution in x is the same for the calibration and reference sample. We also make the assumption that the track kinematics are the same in the calibration and reference samples. Again, this is not realistic but the effects are also true for cases whereby the kinematic distributions disagree!

We then use the bin efficiencies to assign efficiencies to the reference tracks, extracting the adjusted reference track yield l by summing over the bins:

$$l = \sum^{j} \frac{k_j}{\epsilon_j} \tag{3}$$

where k_j is the number of reference tracks in bin j. The ratio of the number of tracks passing PID selection over the adjusted track yield gives the re-weighted efficiency $\bar{\epsilon}$:

$$\bar{\epsilon} = \frac{k}{l} \tag{4}$$

In this example, x is independent of the PID response. Now suppose x is uniformly distributed in both the calibration and reference sample, say between zero and one. By binning in x, we attain the aforementioned Gaussian distribution of bin efficiencies about the integrated effective efficiency. So there are approximately an equal amount of bins with efficiency above the effective value as there are below the effective value. Therefore we expect approximately equal numbers of tracks to be assigned per-track efficiencies in bins which have efficiency fluctuations above the effective value, and equal events assigned those bin efficiencies with values below the effective value.

If we average the calibration dataset over x, all events are assigned the average effective efficiency and the adjusted track yield is then simply

$$l = \frac{k}{\epsilon}. \tag{5}$$

Binning in x, due to the approximately Gaussian distribution of bins there will be an equal number of bins fluctuating upward in efficiency and equal numbers of bins fluctuating down. With a uniform reference sample in x, there will be roughly Gaussian distributed per-track efficiencies assigned to the reference sample. So for one set of reference of tracks k_{j-} falling in bins with efficiency fluctuation y below the average effective efficiency, there is a roughly equal number of tracks k_{j+} falling in bins with efficiency fluctuation y above the average.

Assuming uniform distributions in x there should be equal number of reference tracks falling in the upward and downward fluctuating efficiency bins, so let $k_{j-} = k_{j+} = k_j$. Now:

$$\begin{aligned} l &= \frac{k_j}{\epsilon + y} + \frac{k_j}{\epsilon - y} \\ &= \frac{k_j(\epsilon - y) + k_j(\epsilon + y)}{(\epsilon + y)(\epsilon - y)} \\ &= \frac{2k_j\epsilon}{\epsilon^2 - y^2} \\ &= \frac{2k_j}{\epsilon - \frac{y^2}{\epsilon}} \end{aligned} \tag{6}$$

In the case of zero statistical fluctuations this reduces to the phase-space averaged adjusted yield. For contributions from bins with non-zero y, this always results in an adjusted track yield which is larger. We will therefore always extract a reference $\bar{\epsilon}$ which is always biased downwards.

Even when the x variable is some kinematic variable on which the PID response depends, we expect that some bin efficiencies will fluctuate above their "real" value and some fluctuate down. We have demonstrated that contributions from downward fluctuating efficiencies dominate when the fluctuations, and therefore the binomial uncertainties on the bin efficiencies, are high. This is the downward bias in the re-weighted efficiencies which is purely statistical in nature.

The higher the ratio of the average binomial uncertainty over the average efficiency,

$$\frac{\bar{\sigma_\epsilon}}{\bar{\epsilon}}, \tag{7}$$

the higher the manifested bias becomes. This ratio in typical PIDCalib analyses is tiny and so most do not see this effect, but it is possible that low mis-ID evaluations will be sensitive to it.

3 Toy studies

This is investigated using toy MC, by generating a simulated population of N "candidates" - this is our calibration sample. Each candidate is randomly assigned a value in a control variable, x, which is uniformly distributed between 0 and 1. A random selection is applied to this dataset, with a true efficiency $\epsilon_{\rm true}$. As such, the efficiency of the selection is independent of the control variable x. We calculate the effective efficiency of the selection by counting the number of candidates surviving the selection. We divide the sample into j equal bins in x, and derive an effective efficiency for each bin, ϵ_j.

We generate a second sample of l candidates with the same uniform distribution in x - this is our reference sample. The size of l is ultimately not important, provided the sample is large enough that statistical fluctuations in the reference sample are not important. We use the binned efficiencies constructed using the calibration data to assign per-event efficiencies to the reference sample. We then calculate the adjusted yield for the reference sample, and use this to extract the re-weighted efficiency ϵ. This is, of course, the standard PIDCalib technique. With k_j as the number of reference events in bin j, the adjusted yield l is then:

$$l = \sum^{j} \frac{k_j}{\epsilon_j} \tag{8}$$

and the re-weighted efficiency is then

$$\bar{\epsilon} = \frac{k}{l}. \tag{9}$$

This is carried out for a number of binning schemas for a given calibration dataset, and we track the change in re-weighted efficiency.

3.1 $N = 10^5$, $\epsilon = 0.1$: new proton samples

In this example we have low calibration statistics, so the effects are higher than those in typical hadronic PIDCalib for purposes of illustration. However, this is a similar sample size as that of the new Λ_c based protons samples which are planned to be added to the package. Therefore in the region where these calibration tracks exclusively cover the kinematic space, this gives a demonstration of the expected orders of magnitude of any bias.

The selection efficiency is 0.1, the number of generated calibration events is 100000. The extracted efficiencies are shown in Table 1, along with the typical binomial error for the bin schema, and the evaluation of the bias from the phase-space averaged efficiency.

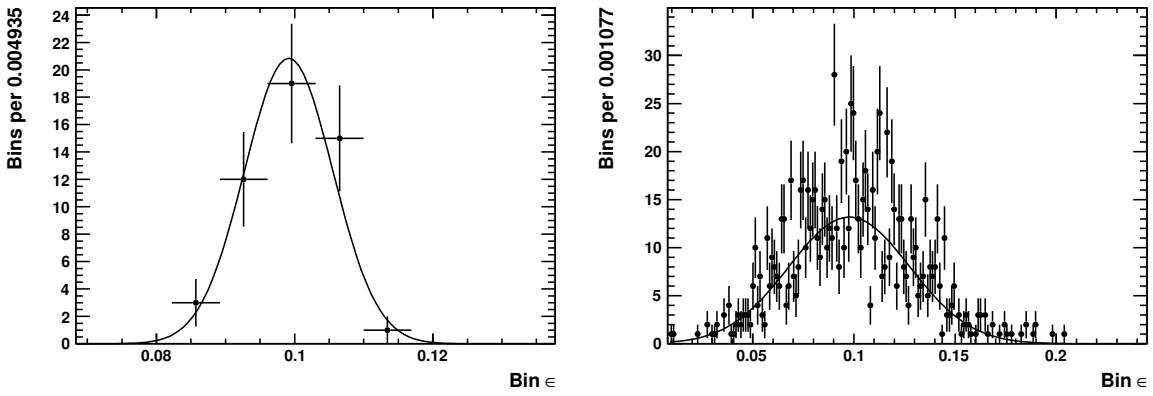
As as the number of bins increases, the fractional binomial efficiency error over the efficiency increases, and the width of the Gaussian distribution of bin efficiencies becomes larger. This leads to larger biases being introduced. Examples of the distributons of bin efficiencies for different numbers of bins are given in Figure 1.

The magnitude of the bias in extracted efficiency is plotted against the number of bins in the re-weighting in Figure 2. The effiency is being clearly biased downward with increasing granularity.

While it is at all times within the typical statistical error on individual bins, this is not indicative of the final statistical error on a re-weighted efficiency. Usually analysts rely on some cancellation on individual bin efficiency errors in a propagated final uncertainty. A more useful comparison may be to the binomial uncertainty on the efficiency of the integrated sample. When we reach 100 bins, a not atypical number of bins in `PIDCalib` schemas, the bias is roughly equal to the statistical error on the integrated sample. As such the bias may be non-negligible.

Table 1: The extracted efficiencies for a variety of binning schemas. In this example, $\epsilon_{\rm true}$ = 0.1, $N_{\rm t}$ = 100000, $\epsilon_{\rm ave} = 9.94 \pm 0.09$ %.

N Bins	$\epsilon_{\rm RW}$ [%]	$\lvert\epsilon_{\rm RW} - \epsilon_{\rm ave}\rvert$[%]	Average bin σ_ϵ[%]	Fractional bias on ϵ [%]
10	9.92	0.01	0.30	0.15
25	9.91	0.03	0.47	0.26
50	9.89	0.05	0.67	0.46
100	9.84	0.10	0.95	0.97
200	9.73	0.21	1.34	2.13
300	9.66	0.28	1.64	2.77
400	9.55	0.39	1.89	3.94
500	9.41	0.53	2.12	5.32
600	9.33	0.61	2.32	6.14
700	9.11	0.83	2.50	8.33
800	9.08	0.86	2.68	8.65
900	8.91	1.02	2.84	10.31
1000	8.80	1.14	2.99	11.45



(a) 50 bins (b) 1000 bins

Figure 1: The bin binomial efficiency distributions distributions for 50 bins (a) and 1000 bins (b).

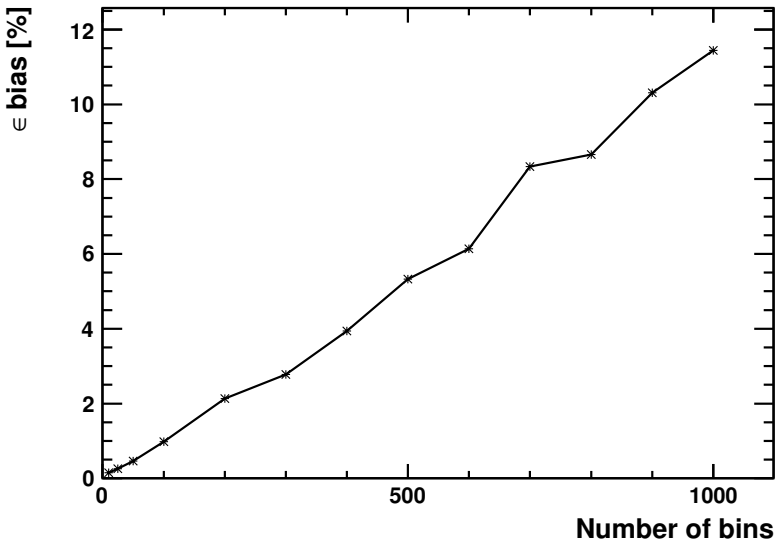


Figure 2: The fractional bias on re-weighted efficiency versus number of bins.

3.2 $N = 10^8$, $\epsilon = 0.002$: $K \to \mu$ mis-ID in tight but realistic cuts

The toy experiment is repeated, but now with a comparable number of calibration candidates as there are calibration tracks for the K species. We take a reasonable estimate for the $K \to \mu$ mis-ID rate as is evaluated for the $B^0_s \to \mu^+\mu^-$ analysis, around 0.002. As such we can estimate the effects of statistical biases in these mis-ID rates.

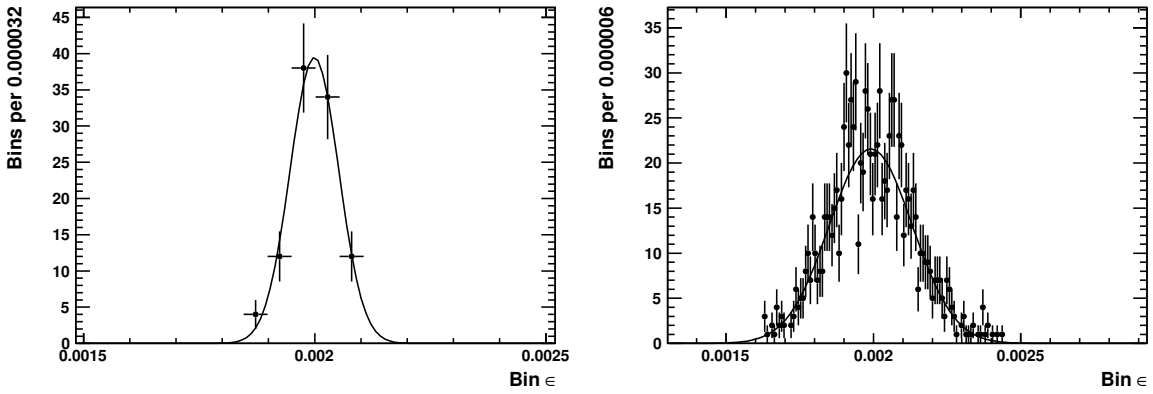
The results are shown in Table 2, again with some example efficiency distributions shown in Figure 3 with a plot of the bias against number of bins shown in Figure 4.

We observe similar forms of bias as for the previous example but much smaller due to the higher calibration statistics. In all cases the bias is within the typical binomial error on the bin efficiencies. In this case, with the higher calibration statistics than our previous example we observe biases which are smaller than the binomial uncertainty on the integrated sample by an order of magnitude. As such any bias in this case should be negligible.

As this effect is a function of the efficiency, for higher efficiency PID cut evaluations with calibration samples of this size, as is the case in typical `PIDCalib` calibrations, the effect is of course negligible.

Table 2: The extracted efficiencies for a variety of binning schemas. In this example, $\epsilon_{\rm true}$ = 0.002, $k = 10^8$, $\epsilon_{\rm ave} = 0.1996 \pm 0.0013$ %.

N Bins	$\epsilon_{\rm RW}$ [%]	$\lvert\epsilon_{\rm RW} - \epsilon_{\rm ave}\rvert$[%]	Average bin σ_ϵ[%]	Fractional bias on ϵ [%]
10	0.19955	0.00000	0.00141	0.00201
25	0.19952	0.00003	0.00223	0.01576
50	0.19949	0.00006	0.00316	0.03108
100	0.19944	0.00011	0.00446	0.05677
200	0.19937	0.00018	0.00631	0.09012
300	0.19929	0.00026	0.00773	0.13221
400	0.19920	0.00035	0.00893	0.17534
500	0.19906	0.00049	0.00998	0.24695
600	0.19896	0.00059	0.01093	0.29331
700	0.19886	0.00069	0.01181	0.34374
800	0.19877	0.00078	0.01262	0.39152
900	0.19866	0.00089	0.01339	0.44389
1000	0.19859	0.00096	0.01411	0.48129



(a) 100 bins (b) 1000 bins

Figure 3: The bin binomial efficiency distributions distributions for 100 bins (a) and 1000 bins (b).

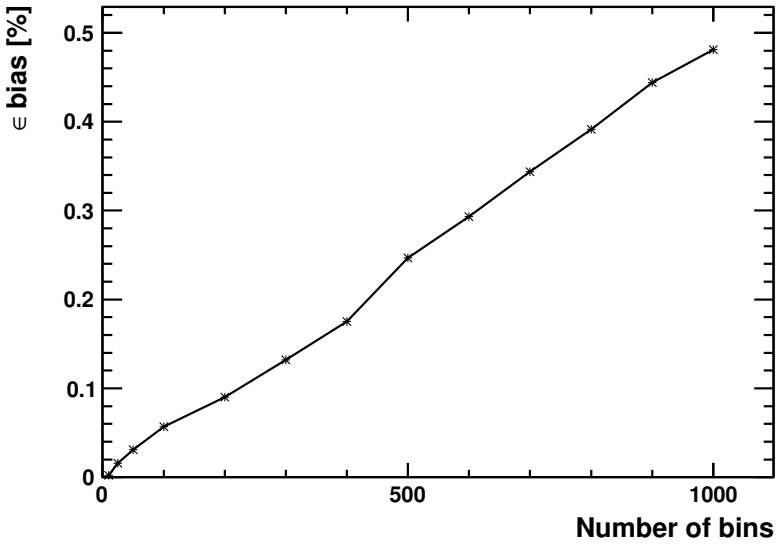


Figure 4: The fractional bias on re-weighted efficiency versus number of bins.

4 Conclusion

A purely statistical bias which arises in efficiency re-weighting procedures was identified and explained. The possible effects of this bias on `PIDCalib` analyses was investigated using toy MC.

It was found that for most high statistics applications, for example in the $K \to \mu$ mis-ID estimations, the bias is negligible. However, in the instances where low statistics are available, as is the case for the new proton samples, this effect is shown to be at the level of the efficiency binomial error on the integrated samples, and becomes significant.

It should also be stressed that this is an effect which arises via the re-weighting procedure and affects the average re-weighted efficiencies of reference samples - it cannot explain why `PIDCalib` has in some cases been shown to overestimate/underestimate the mis-ID efficiencies in individual kinematic bins in some very low efficiency cases.